图书情报工作
第 63 卷 第 16 期 2019 年 8 月

中文超声文本结构化与知识网络构建方法研究*

■ 尚小溥[1] 许吴环[1] 赵红梅[1,2] 张润彤[1] 朱槃[1]
[1] 北京交通大学经济管理学院信息管理系 北京 100044 [2] 北京大学人民医院 北京 100044

摘要：[目的/意义]超声检查是判断患者病情的重要依据，目前主要检查数据是以文本形式存在。本文提出一种基于超声检查数据的文本结构化和知识网络构建方法，为进一步挖掘临床知识奠定数据基础。[方法/过程]对自然语言处理技术在超声文本环境下的应用进行改进，包括分词处理、内容定位、结构化识别三个主要步骤，实现对超声文本的切分与标记，并且在此基础上建立其结构化知识网络。[结果/结论]真实数据测试结果显示，本文提出的面向超声检查文本的结构化方法具有较好的性能表现。该方法可以实现对批量超声文本结构化网络的自动构建，能够反映超声文本中结构化内容的层次关系与属性结构等潜在知识。

关键词：超声文本 自然语言处理 文本结构化 知识网络
分类号：TP393
DOI：10.13266/j.issn.0252-3116.2019.16.012

电子病历作为一种专业的治疗过程全记录载体，是当前医疗实践中最重要的文档资料，也是临床实践的知识库[1]。电子病历中的专业医学检验检查数据，是医疗大数据分析中重要的客观数据资源[2]，也是循证医学中的重要数据支撑。超声检查作为临床医学中重要的检查手段，具有快速直观判断特定部位病情的特点。然而，与大多数的医学影像检查以及常规的检验化验不同，超声检查结果在数据的呈现形式上仅表现为医生录入的文本数据。文本数据作为一种非结构化数据，一直是实现精准计算机数据分析、知识挖掘等工作中需要解决的重要问题。因此，超声检查数据的结构化以及结构化数据的知识网络构建是医疗大数据分析和临床医学研究中亟待解决的重要问题。

作为一种专业的医学文本数据，超声检查数据与一般的日常自然语言和文本数据相比，呈现出以下独有特点：总体语言风格与日常用语差异较大，专业词汇较多且存在异形词。上述问题给超声检查数据的结构化与知识网络构建提出了巨大挑战。传统自然语言分析与结构化处理方法在此场景下直接应用，无法得到较高精度的结果，难以满足相应医疗大数据研究的进一步分析等工作。本文根据超声数据的特点，基于自然语言处理技术（Natural Language Processing, NLP），针对性地提出一种文本分析与结构化的系统方法，该方法能够实现对超声文本数据较高精度的分解与标注，具备自动构建知识网络的能力，具有重要的科学意义和应用价值。

1 研究现状

自然语言处理一直是文本挖掘领域中的基础性问题。在海量文本中挖掘出隐藏的知识一般可以从两种层面实现：一是仅对特定信息的搜索与抽取，根据抽取到的信息进行进一步的知识挖掘[3]；二是对文本进行全面结构化，将全部内容均转化为能够被计算机识别的单词，进而对这些结构化的单词进行关系网络的建立，形成知识图谱，从而通过推理等方式实现对知识的挖掘[4]。前者在挖掘对象明确、对要挖掘知识具有逻辑性认知的场景中效率较高，但对于未知知识，或所挖掘的知识并不具备较强的目标性，则需要对文本进行全面结构化分解，采用第二类知识挖掘方法。其次，中文文本与英文文本在处理过程中存在一定差别：英文

* 本文系国家自然科学基金项目“面向临床决策辅助的电子病历文本结构化方法与知识挖掘研究”（项目编号：61702033）和教育部人文社科项目“基于电子病历文本的临床知识挖掘研究”（项目编号：17YJC870015）研究成果之一。

作者简介：尚小溥（ORCID：0000-0002-7872-5744），讲师，博士，E-mail：sxp@bjtu.edu.cn；许吴环（ORCID：0000-0003-2621-7913），硕士研究生；赵红梅（ORCID：0000-0001-6880-3342），副研究员，硕士；张润彤（ORCID：0000-0003-0246-5058），系主任，教授，博士；朱槃（ORCID：0000-0002-5802-8132），本科。

收稿日期：2018-11-20 修回日期：2019-03-22 本文起止页码：112-120 本文责任编辑：杜杏叶

尚小溥,许吴环,赵红梅,等.中文超声文本结构化与知识网络构建方法研究[J].图书情报工作,2019,63(16):112-120.

文本是以单词为单位,单词间采用空格符分开;中文文本是以字符为单位,字符间没有分隔符。在此背景下,一些研究专门关注如何对中文文本进行分词[5],并在此基础上建立相应领域具有标注的词典[6],该词典可以成为对同类中文文本分词的重要依据。当前也有一些分词工具支持中文文本切分,比如 Stanford NLP[7]、Jieba[8]、哈工大 LTP[9] 等,这些分词工具对于一般日常文本能取得一定效果,但对专业的医疗文本来说效果欠佳,无法满足对文本数据进行进一步分析的技术要求。

当前已有一些专门面向医学领域场景、基于医学相关文本信息与知识挖掘方法的研究,如基于人工建立的语料库,针对医学领域的学术文献进行中英文的文本分析与对比[10];针对医学学术文献研究引文上下文内容的信息价值问题[11];有依赖于字符包、单词包、字符嵌入和单词嵌入,构建出一个基于序列标记的中文临床笔记推测检测系统,并证明了分词在中文临床自然语言处理中的重要性[12];有研究通过聚类的方法,对医学文献进行挖掘,用于分析近年来的研究方向和热点[13];有通过共现计数分析临床变量之间的潜在依赖性,以促进改进的高维倾向评分特征选择的发展[14];还有针对互联网医学信息资源,提出了一种基于细粒度语义化描述的医学文本检索算法,在检索结果的选择方面,采用相似度计算方法实现对相关内容的匹配[15];也有专门关注互联网医疗相关文档中的语义识别问题[16]的研究。然而,上述研究均是针对公开的医学学术文献或专业资料进行的文本数据分析,且部分研究基于英文文献进行分析,从技术实现角度来看,这与中文临床病历文本分析挖掘有一定的区别。

电子病历文本是一种重要的医疗文档,记录了临床诊疗过程中的各种检查、病情、诊断等信息。近年来国内开始有学者关注对电子病历的文本挖掘工作。专门针对中文电子病历文本,研究了在利用既有分词工具基础上的分词方法,其精度最高可达 78.06%[17];有研究以电子病历文本为基础,挖掘出院记录部分潜在语义[18],但该研究只针对四种治疗方案进行了评估,评估结果粒度较大,应用于临床实践的针对性不强;还有一些研究基于电子病历开展临床决策支持的相关探索[19-21]。这些研究的重点多在电子病历中的结构化和半结构化数据,或是较有针对性地抽取特定关键词等信息[22]。除此之外,已有少量针对非结构化医疗文本的研究,H. Wang 等利用自然语言处理方法从中文肝癌手术记录中提取了肿瘤相关信息的发展和评估[23];B. He 等从语法和语义的角度提出几种中文临床文本的注释方法,并构建了基于 NLP 模块的综合语料库,但该语料库覆盖率和注释效率均较低[24]。

将电子病历自由文本转化为计算机可处理识别的规律形式,是知识挖掘工作的前提,往往需要多种自然语言处理技术综合运用。当前中文医学文本结构化的常见方式是将数据转化为 <指标:指标值> 的形式,主要是以人工构建的指标词库为依据,通过信息抽取的方式,从非结构化的文本中抽取出特定形式[25-27]。实体识别是文本结构化中的重要目标,有研究通过深度学习的方式为电子病历文本的实体属性赋予标签[28],以及疾病名称[29]、医疗事件名称[30]的识别等。然而上述工作并没有关注实体间的关系,但若是应用于临床决策支持或临床数据分析的场景,往往需要能够客观反映数据间的逻辑关联。有研究关注病历中的实体与实体间的关系,探索自动构建英文电子病历文本的知识图谱的方法[31-32]。当前,也出现了尝试中文电子病历知识图谱构建的研究[33]。然而中文电子病历知识图谱构建的研究刚刚起步,且自动化程度低。基于电子病历的知识图谱是临床知识推理、诊断的可靠基础,图谱中节点的关系是一些特定的语义关系,如检查-疾病,疾病-症状的关系等。超声检查文本是电子病历的重要组成,是对患者所检部位超声影像的详细描述,上述的特定关系并不存在。因此,本文关注超声文本中的网络结构,刻画超声文本中"实体-属性-值"间的连接关系,以及实体间的层次关系,有机构成了超声知识网络。

本文提出一套自动化处理流程,通过分词处理、内容定位、结构化识别三个主要步骤,实现对超声检查文本的全面结构化。该结构化网络在充分保留电子病历信息的同时,为各类数据分析需求奠定最客观的数据基础,进一步推动相关医学研究和临床护理。

2 超声文本数据的结构化与知识网络构建方法

2.1 总体流程

本文提出的超声检查文本结构化与知识网络构建方法,主要由分词处理、内容定位、结构化识别三个主要步骤组成。分词处理阶段基于对超声文本分词特点,提出分词矫正算法;内容定位经过文本聚类,与文本间相似短句定位映射,实现超声文本相同语义内容的归类映射;结构化识别阶段,基于前述处理提出一种实体属性值识别算法,并根据识别结果,将超声文本映射到网络结构。将该方法输入的是批量的超声检查自由文本,输出的是结构化后的数据,可以存储在关系型

数据库中。图 1 给出了方法的总体步骤，每个步骤的具体实现过程在后续小节中详细阐述：

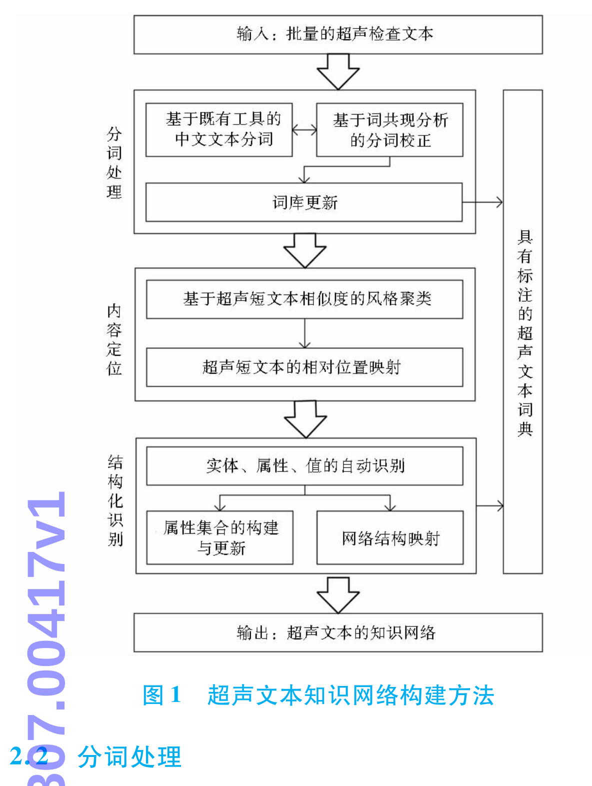


图 1　超声文本知识网络构建方法

2.2　分词处理

上文已经提到一些使用较广泛且能够支持中文分词处理的自然语言处理工具[7-9]，其中 Stanford NLP 和 Jieba 是开源工具，本研究采用 Stanford NLP 来进行初步的文本切分处理。在处理日常语言方面，Stanford NLP 具有较高的性能，但是对于相对专业的医学文本，其处理能力欠佳，而高质量的分词结果对临床 NLP 任务是至关重要的[12]。一种可行的方式是增加专业词库，在通过既有 NLP 工具处理的基础上，采用基于词共现分析的方法，实现专业词库的自动补充。

对于 Standford NLP 等相关分词工具，一般情况下，若出现了词库中没有的新词，会通过特定算法实现未登录词的切分。对于非理想状态的切分结果，存在三种情况：一是分词算法将本可以合并在一起的字/词切分开，相应的切分结果无法将可固定搭配的词/词组呈现出来，本文称为“过切分”；另一种情形是分词算法把本应该切分为两个或者更多的字/词，判断为一个词/词组的内容，本文称为“欠切分”；第三种情形是分词工具在不恰当的位置进行了切分，将原本应该在一起的词分开，而不应该在一起的词切分在了一起，本文称为“误切分”。由于超声文本中存在大量的缩略词和特殊名词等，对其分词主要是“过切分”（见表 1（a））和“误切分”（见表 1（b））的问题。对切分结果的判断，应该考察切分后的词/词组其是否正确表达了文本的含义。

表 1　非理想状态的文本切分结果举例

（a）过切分	（b）误切分
未 | 见 | 明显 | 异常（斯坦福 NLP）	腔 | 内强 | 回声（斯坦福 NLP）
未见 | 明显异常（理想切分）	腔内 | 强回声（理想切分）

对于过切分的处理，本文在研究中采用了基于词共现分析的分词校正方法，具体通过对既有分词工具得到的初步结果进行相邻词的共现分析，识别和判断非理想情况的切分，并实现对切分结果的校正优化。同时，针对“过切分”情况识别出的新词，也可以对“误切分”情况带来一定的改善。如，存在“强回声”被过切分为“强”+“回声”，当“强回声”被正确识别时，表 1（b）中的误切分也可被改善。

本文采取的词共现频率计算方式如下：

令 $S=\{W_1, W_2, \cdots, W_n\}$，$S$ 代表某条数据记录，W_i 表示该记录的第 i 个词。W_i 在文本中出现的次数记作词频 Cnt。

定义 1. 词对 (w_i, w_{i+1}) 的右共现频率定义为 $F_{R(w_i,w_{i+1})} = Cnt(w_i, w_{i+1}) / \sum_{\chi\in A} Cnt(w_i, \chi)$，其中，$A$ 是文本中所有位于 w_i 右边的词的集合。

定义 2. 词对 (w_i, w_{i+1}) 的左共现频率定义为 $F_{L(w_i,w_{i+1})} = Cnt(w_i, w_{i+1}) / \sum_{\chi\in B} Cnt(\chi, w_{i+1})$，其中，$B$ 是文本中所有位于词 w_{i+1} 左边的词的集合。

算法 1：基于词共现分析的分词校正算法的核心伪代码如下：

输入：文本中的相邻词对
输出：候选新词词典 *Dic*

```
for (w_i, w_{i+1}) do
    if Σ_{χ∈A} Cnt(w_i, χ) > 1 then
        fre_R ← Cnt_{w_i,w_{i+1}} / Σ_{χ∈A} Cnt(w_i, χ)
        if fre_R > = C then
            dic.append(w_i, w_{i+1})
    if Σ_{χ∈B} Cnt(χ, w_{i+1}) > 1 then
        fre_L ← Cnt_{(w_i,w_{i+1})} / Σ_{χ∈B} Cnt(χ, w_{i+1})
        if fre_L > = C then
            dic.append(w_i w_{i+1})
Delete repeated words in Dic
Word segmentation again with Dic
```

针对实验所采用的数据及实验分析，将阈值 C 设置为 0.9，即右共现频率或左共现频率大于等于 0.9 的

组合词“W_iW_{i+1}”为候选新词。根据实验结果分析，阈值为0.9时能过滤绝大多数干扰项，同时保留较多的新词。最终得到基于超声文本的领域词典 *Dic*。

算法1中输入的词对(w_i, w_{i+1})满足如下规则：若两个词由标点符号隔开，则不做共现分析；第2和第6步中，设置某个词出现次数大于1才做共现统计，是因为只出现一次的词根据定义1和2计算的共现频率一定为1，然而这些词对并不符合本文发现新词的思想，且绝大多数均为干扰项，故过滤。对于一个专业术语被切分成三个或四个词的情况，本文通过迭代上述分词算法进行处理。例如，“肝内外胆管”初始被分为“肝”+“内外”+“胆管”，第一次分词矫正时得到组合词“肝内外”，加入词典 *Dic* 后，第二次被分为“肝内外”+“胆管”，第二次可发现新词“肝内外胆管”。在实验分析中发现，超声检查文本中一个专业术语最多被切分成不超过四个词，且第三次执行时可发现的新词数量已经很少，故迭代次数设置为3。

2.3 内容定位

2.3.1 文本聚类 由于文本聚类依赖于文本之间的相似度，所以需要计算每两个文本之间的相似度，从而得到文本相似度矩阵，进而利用相似度矩阵实现文本聚类，达到提高后续实体、属性、值的识别能力，提升识别精度的目的。

医学文本采用的都是相对专业和直接的表述方式，尽管医生可能使用不同的词汇描述同一种情形，但基本不存在一词多义的情况，因此文字层面的相似度即可评估内容的相似程度。本文采用海明距离来评估每一例超声报告与其它报告的相似程度，并根据不同的相似度对这些超声文本进行聚类，即同一类的超声文本具有较高语言相似度，而不同类别的超声检查文本之间的相似度较低。又因为每条超声检查文本记录较长(200～300字)，且数据量较大，所以本文采用SimHash算法的降维思想[34－35]，再将得到的相似度矩阵通过谱聚类算法聚为 K 类。

算法2：超声文本聚类算法的伪代码如下：

```
for each Record do
    s_i ← Finger Print (Record)
for each s_i do
    d(s_1, s_2) ← HammingDistance(s_i, s_j)
    sim(s_i, s_j) = 1 − d(s_i, s_j)/hashBits
    M.append(sim)
SpectralClustring(M, K)
```

聚类数目参数 K 分别设置为3、4、5、6、7，根据对4 000条数据的分析观察，将文本分为五类时能有效将电子病历中的最不相似的电子病历分开，且便于后续实验进行，于是本文在对实验数据处理时决定选用 $K=5$ 的聚类方案。

2.3.2 相似短句定位映射 在对超声文本进行了相似度聚类分析的基础上，进而实现各类超声文本中短句相对位置的定位与映射。如记录a与记录b中有数量相近的若干短句，本文试图建立a[x]与b[y]之间的映射关系，映射目的在于识别出不同超声文本中对同一现象的描述部分。本文同样采用了上述海明距离来评估不同超声报告短句之间的相似程度。

算法3：超声文本中的短文本相对位置映射

①对每条记录以标点符号为界进行短句切分；

②选择包含短句数目最多的记录作为第一条记录；

③计算第 i 条记录中的第j个短句与第一条记录中第 m 个短句的相似度 $sim(s_{1_m}, s_{i_j})$，其中 $i=2,3,\cdots,n$；

④提取第 i 条记录中的与第一条记录中第 m 个短句相似度最高的短句；

⑤对第一条记录中的所有短句做相同处理，得到相似短句映射表。

基于上述算法，将每一例超声检查文本间相似度最高的短句，作为相对位置匹配的一组短句，得到相似短句映射表。短句相对位置的映射，其实是对不同超声检查病例中描述相同语义内容的短句的识别与定位，为后续实体、属性、值的识别奠定基础。

2.4 结构化识别

在实现了对超声检查文本内容定位的基础上，可进一步通过提出的算法实现对切分内容进行“实体、属性、值”的标记，从而建立起具有层次结构的超声知识网络。

2.4.1 实体、属性、值的识别 实体和属性作为相对客观的描述对象，在超声检查文本中其用词一般相对固定。值作为实体和属性的具体定量或定性内容，其往往呈现出较为丰富的内容。且由于中文的书写习惯，“值”通常出现在短句的末尾，表现为数字或文字形式，然而在分析中发现，超声文本存在一些“汉字值”出现在属性之前，如“类圆形/无回声”。据此，本研究根据具有映射标记的短句组内固定词语与相对变化词语的规律特征，识别实体、属性以及属性值。

算法 4:实体、属性、值的识别

①统计组内短句频数,选取出现次数最多的短句作为标准句;

②选取 $sim(s,s_i)>0.5$ 的短句构成一个集合;

③对集合中的短句进行分词;

④分别以每一个短句的分词结果为基础与其后面的短句分词结果作比较,如后者为前者的子集,则将后者删除;

⑤对④的结果集合 S 中的短句分词,统计每个词的频数 $Cnt(w_i)$ 及其相对于短句总数的频率 f,其中 $Cnt(w_i)$ 最大且 $f\geqslant 0.8$ 的词可认为是实体;

⑥在每个短句找到实体出现的位置 o,如果存在判断 $o+1$ 若是最末尾位置 ,则执行(7),否则执行(8);

⑦判断 $Cnt(w_{o+1})\leqslant P$,若是则为值,否则是属性;

⑧判断 $Cnt(w_{o+1})>P$ 或($Cnt(w_{o+2})>Q$ and $o+2\neq e$),若是则为属性,否则为值。

根据本文选用的实验数据,这里参数 P 和 Q 设置为:$P=Q=$(S 中包含实体的短句数目)/2,此时取得较好的识别效果。

根据相似短句定位映射的结果,对每一组相似短句进行实体抽取。首先在组内选取重复出现次数最多的短句作为实体抽取的标准。$sim(s,s_i)>0.5$ 是为了过滤垃圾数据。在 2.2 节自定义词典 *Dic* 的基础上,对该集合中的每一个短句进行分词,并依次以其分词结果为基础,与其后面的短句分词结果作比较,若后者为前者的子集则认为两个短句的描述一致,将后者短句删除。例如,集合中有下列短句:“肝脏大小形态可”“肝脏形态可”“肝脏形态大小可”“肝脏形态饱满”“肝脏形态失常”“肝脏形态略饱满”“肝脏失常态”,它们的分词结果记为 A{肝脏,大小,形态,可},B{肝脏,形态,可},C{肝脏,形态,大小,可}。B 和 C 均为 A 的子集,所以将“肝脏形态可”“脏形态大小可”从集合中移除。所以集合变为{肝脏大小形态可,肝脏形态饱满,肝脏形态失常,肝脏形态略饱满,肝脏失常态}。对集合中剩下的短句分词,并统计 $Cnt(w_i)$ 与 f,其中 $Cnt(w_i)$ 最大且 $f\geqslant 0.8$ 的词认为是实体。识别出实体后根据算法 4 的(6)(7)(8)进行属性和值的识别。如在上面例子中,分词后出现次数最多的词为“肝脏”且频率为 1.0,所以“肝脏”记为实体。在第一句中,w_{o+1} 为“大小”,接着看 w_{o+2}“形态”,$Cnt(w_{o+2})>Q$ 且 $o+2\neq e$,所以“大小”“形态”记为属性,“可”记为值。其余短句同理,最终可得到实体抽取结果示例如表 2 所示:

表 2　实体抽取结果示例

实体	属性	值
肝脏	大小、形态	可、饱满、略饱满、失常、失常态

2.4.2　网络结构映射　实体、属性、值识别之后的词也就构成有了相应标记的实体库、属性库与属性值库。对于腹部超声,一般情况需要检查:肝、胆、胰、脾、肾五个器官。据此场景,本文以关键词为依据,将上述器官作为超声文本描述对象的分隔符,从而实现对不同器官描述的平行关系区分,在同一个器官描述部分,相似映射的短句为平行关系。将待处理短句的分词结果映射到上述识别结果库,得到带有识别标记的词,根据标记组织成目标结构化形式。本文提出结构化存储的一般形式为:(一级实体[,二级实体][,属性][,属性值]),这种形式同时刻画了短句内“实体 - 属性 - 值”间的连接关系,与短句间实体的层次关系。其中,一级实体主要为上述固定的五个检查器官;“属性”或“属性值”可能出现为空的情况。以 2.4.1 节对算法 4 举例说明的短句为例,根据表 2 的识别结果,其部分结构化存储记录为“肝脏 - 大小 - 可”,“肝脏 - 形态 - 可”,“肝脏 - 形态 - 饱满”等。本文提出的方法流程中,超声文本的结构化处理是建立相应超声知识网络的基础,每一条结构化存储记录,都是知识网络中的一条路径。最终,可以通过可视化工具(如 D3. js),将上述形式存储的超声知识网络结构展现出来。

3　数据实例测试与分析

基于本文提出的超声检查文本结构化方法,本章采用真实数据对算法的实现过程进一步进行阐述与验证。

3.1　数据来源与测试方法

本研究的数据来源于某大型三甲医院超声科的腹部超声检查数据,总数据条数为 4 818 条。数据在使用前经过了脱敏处理,隐去了能够识别出患者的相关信息,包括患者姓名、患者 ID、就诊时间等内容。只保留了“超声所见”字段。研究随机选取了其中 4 600 条数据进行训练,其余 218 条数据进行测试,训练数据与测试数据不存在交集。测试数据同时经过人工分词与实体标记,相关性能表现通过人工标记结果与本文所提出方法运行得到结果进行对比得到。

3.2　分词处理效果分析

对上述 218 份随机测试数据采用经 2.2 节得到的

领域词典进行分词调整。图 2 展示了采用 Stanford NLP 预处理结果与基于词共现分词调整结果的对比。

肝脏/:/形态/异常/,/肝/右/叶/最大/斜/径/10.2/Cm/,/
左/肝/上/下径/4.7/Cm/,/包膜/欠/光滑/,/边缘/钝/,/
实质/回声/增强/增/粗/,/分布/欠/均匀/,/
肝/内/血管/走形/欠/清晰/,/肝/内/胆管/无/扩张/。/

图 2(a) 采用 Stanford NLP 的分词结果

肝脏/:/形态/异常/,/肝右叶/最大斜径/10.2/Cm/,/
左肝/上下径/4.7/Cm/,/包膜/欠/光滑/,/边缘/钝/,/
实质回声/增强/增粗/,/分布/欠均匀/,/
肝内血管/走形/欠/清晰/,/肝内胆管/无/扩张/。/

图 2(b) 基于词共现分析分词矫正的分词结果

进一步,基于上述分词结果,以人工分词结果为标准,对本文方法与通用分词工具的结果进行分析,得到了图 3 所示的准确率、召回率和 F1 指标,计算公式如下。从这些指标可见,基于词共现分析的方法得到专业词典,有效提升了分词结果的精度。

准确率 = 正确分词的数目/分词总数目 * 100%

召回率 = 正确分词的数目/标准分词总数目 * 100%

F 值 = 准确率 * 召回率 * 2/(准确率 + 召回率) * 100%

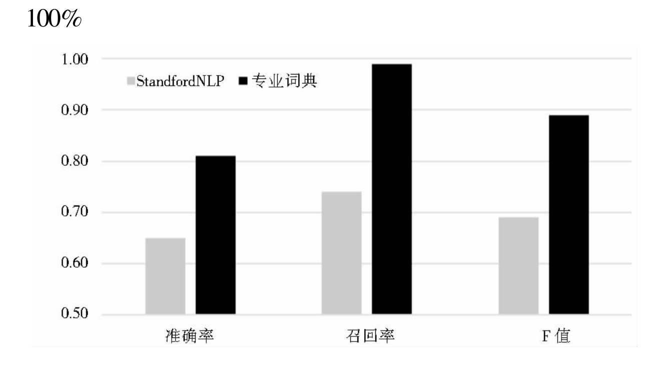


图 3 分词效果对比

3.3 内容定位结果与展示

在分词的基础上,对测试数据进行内容定位。首先根据算法 2,对实验文本进行聚类。图 4 展示了 50 例超声检查文本的相似度结果,图中第 i 行 j 列的方块代表第 i 条超声检查文本和第 j 条超声检查文本之间的相似度,颜色越深表明相似程度越高。

图 4 超声检查文本的相似度热力图

进一步在不同类别的文本分类中，进行短句定位，即算法 3。图 5 给出了在聚类基础上内容定位映射的部分结果：

0	1	2	3	4	5	6
肝脏大小形态可	表面平滑	边缘不钝	肝右叶无回声	肝内血管结构显示清晰	门脉不宽	血流通畅
肝脏大小形态可	表面平滑	边缘圆钝	实质回声弥漫不均	肝内血管结构显示清晰	门脉宽	壁不厚
肝脏大小形态可	表面平滑	边缘不钝	实质回声均匀	肝内血管结构显示清晰	门脉不宽	血流通畅
肝脏大小形态可	表面平滑	边缘不钝	肝右叶无回声	大小	边界清	血流通畅
肝脏形态大小可	表面平滑	边缘不钝	肝内限局性低回声	大小	门脉不宽	血流通畅
肝脏形态大小可	表面平滑	边缘不钝	实质回声欠均匀	右叶见无回声	门脉宽	血流通畅
肝脏形态大小可	表面平滑	边缘略钝	实质回声弥漫增粗增密不均	肝内血管结构显示欠清晰	门脉宽	血流通畅
肝脏大小形态可	表面平滑	实质回声弥漫增粗	实质回声弥漫增粗	肝内血管结构显示清晰	门脉不宽	血流通畅
肝脏大小形态可	表面平滑	边缘不钝	实质回声增粗不均	肝内血管结构显示清晰	大小	血流通畅

图 5　内容定位部分结果

3.4　结构化识别结果分析

在进行了内容定位的基础上，本文进一步根据算法 4 测试了其结构化识别能力。识别结果分别被标记为实体、属性、值三类。从测试文本中随机抽取 10、50、100、150、218 份超声检查文本记录，进行结构化识别后实体、属性及属性值识别的准确率如图 6 所示：

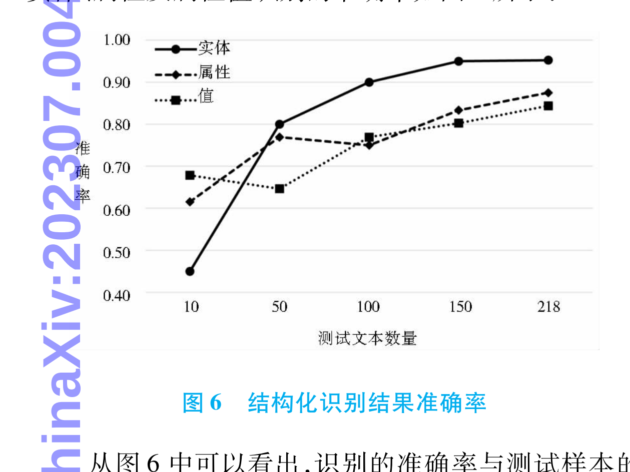


图 6　结构化识别结果准确率

从图 6 中可以看出，识别的准确率与测试样本的数量相关，在一定范围内，总体呈随样本数量增加而上升的趋势。医生在书写超声检查文本时遵循一定的书写规范，实体和属性是比较有限的检查对象，在不同的病人记录中会反复出现，而其对应的取值则更丰富多变。本文提出的结构化识别思想正是基于这种规律，所以当文本数量大时上述规律则体现的更明显。不同对象的检查实体相对来说较固定，所以表现出较好的识别效果，而属性和值的情况更为复杂。

3.5　超声检查知识网络可视化

本文对训练数据在上述结构化识别的基础上，通过确定实体间的层次关系，建立了如图 7 所示的超声检查知识网络，该网络充分保留了超声检查知识，可结构化存储，为更高层次的智能诊断决策应用场景提供基础。

4　总结与展望

本文提出了一种面向超声检查文本的结构化与知识网络构建系统方法，该方法是一套具有创新性的整体流程，通过对多种算法的综合运用，实现对批量医疗检查文本的自动结构化、自动构建网络关系，可为电子病历结构化研究提供一个新思路。在分词处理阶段，通过对相邻词的共现现象分析，更新并建立了领域词典，用该方法进行分词纠正后，准确率相比现有分词工具提高了 16%。内容定位从检查记录和组成记录的短句两个层次上，根据文本相似度对相同检查对象的描述部位进行分组，以提高结构化识别的精度。通过对真实数据的测试与分析发现，本文实体、属性和属性值的识别算法准确度随着样本数量的增多，总体上呈上升趋势，且对较大批量的数据表现出了较好的识别效果。

本文的研究也存在以下不足：①不适用于小量数据，由于在样本数量较小时，描述同一实体的一组相似短句，其属性和值相对固定与变动的规律不易体现，易导致算法 4 错误识别，这是本文算法的局限性；②算法 4 中的参数 P 和 Q，针对不同的实验文本可能需要调整和训练。改进上述问题是我们今后的工作方向。

今后工作可以在本文所提出相关方法的基础上，研究对更多类型医学文本的结构化与知识网络构建，从单一类型医疗文本数据的知识网络构建，发展成为全景式的医疗文本的结构化与知识网络构建，为充分挖掘医疗文本中隐藏的知识奠定数据治理基础。

尚小溥,许吴环,赵红梅,等.中文超声文本结构化与知识网络构建方法研究[J].图书情报工作,2019,63(16):112－120.

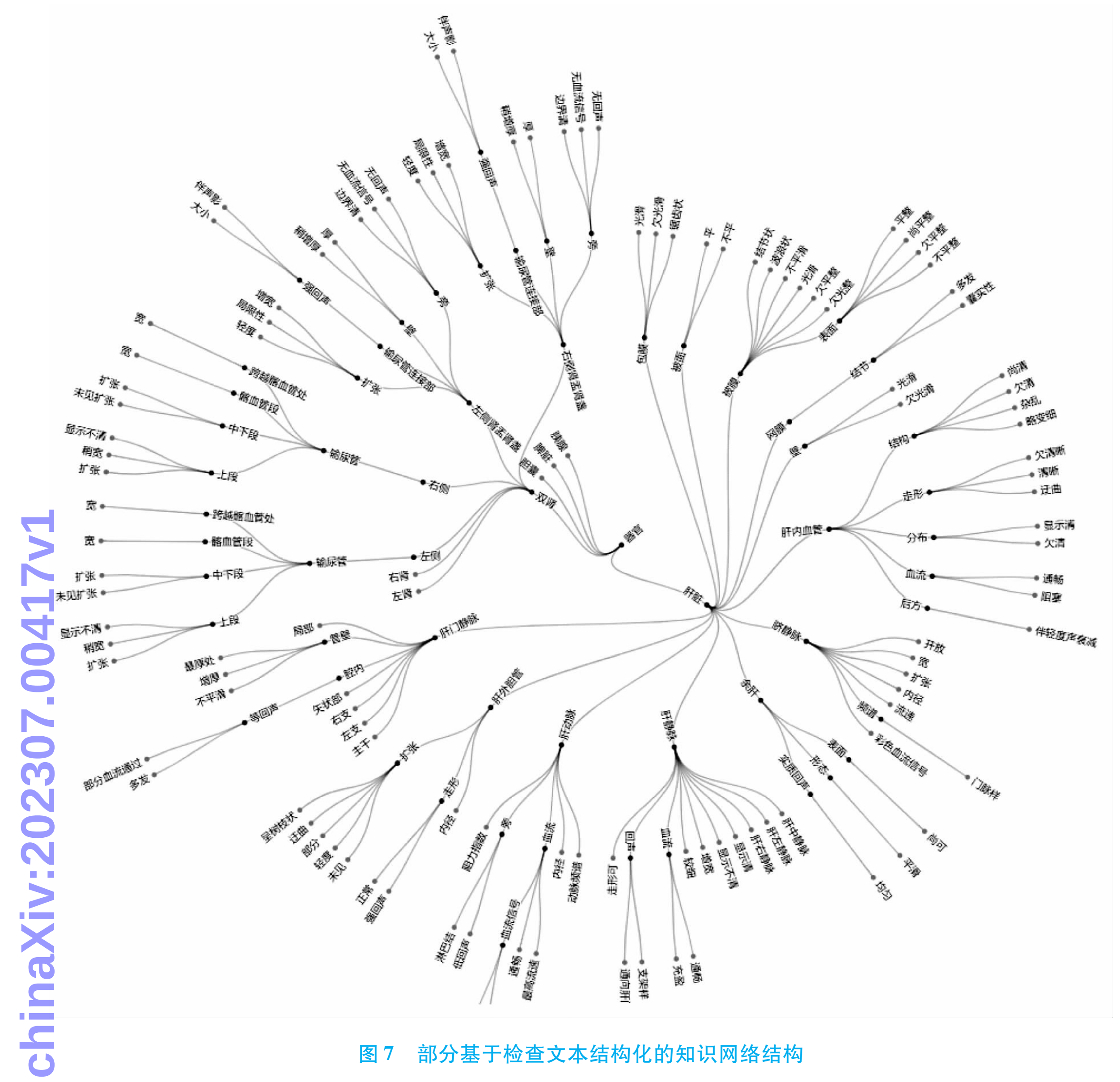


图7 部分基于检查文本结构化的知识网络结构

参考文献:

[1] 陈永莉,洪漪.检索语言在医学信息管理与检索中的应用综述[J].图书情报知识,2015(3):72－79.

[2] 郭熙铜,张晓飞,刘笑笑,等.数据驱动的电子健康服务管理研究:挑战与展望[J].管理科学,2017,30(1):3－14.

[3] JIMÉNEZ P, CORCHUELO R. On learning web information extraction rules with TANGO[J]. Information systems, 2016, 62(12): 74－103.

[4] 刘峤,李杨,段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3):582－600.

[5] 张义,李治江.基于高斯词长特征的中文分词方法[J].中文信息学报,2016,30(5):89－93.

[6] 郭顺利,张向先.面向中文图书评论的情感词典构建方法研究[J].现代图书情报技术,2016,32(2):67－74.

[7] STANFORD NLP. The stanford natural language progressing group [EB/OL]. [2018－06－09]. https://nlp.stanford.edu/.

[8] JIEBA.结巴中文分词[EB/OL].[2018－04－09]. http://www.oss.io/p/fxsjy/jieba.

[9] LTP.语言云[EB/OL].[2018－04－08]. https://www.ltp-cloud.com/.

[10] 王兰英,雍文明,王连柱,等.中美医学论文英文摘要文本对比分析[J].科技与出版,2011(11):78－82.

[11] 刘洋,崔雷.引文上下文在文献内容分析中的信息价值研究[J].图书情报工作,2014,58(6):101－104.

[12] ZHANG S, TIAN K, ZHANG X, et al. Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models[J]. Journal of biomedical informatics, 2016, 60: 334－341.

[13] 于跃,徐志健,王坤,等.基于双聚类方法的生物医学信息学文本数据挖掘研究[J].图书情报工作,2012,56(18):133－136.

[14] FINLAYSON S G, LEPENDU P, SHAH N H. Building the graph of medicine from millions of clinical narratives[J]. Scientific data, 2014, 1: 140032.

[15] 郭少友,李亚菲,梁园园.基于细粒度语义化描述的医学文本检索[J].情报理论与实践,2015,38(8):130－134.

[16] 魏巍,郑杜. 融合统计学习和语义过滤的 ADR 信号抽取模型构建研究[J]. 图书情报工作,2017,62(5):115 - 124.
[17] 李国垒,陈先来,夏冬,等. 中文病历文本分词方法研究[J]. 中国生物医学工程学报,2016,35(4):477 - 481.
[18] 张晔,张晗,尹玢璨,等. 基于电子病历利用支持向量机构建疾病预测模型——以重度急性胰腺炎早期预警为例[J]. 现代图书情报技术,2016,32(2):83 - 89.
[19] LEI J, TANG B, LU X, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American medical informatics association, 2014, 21(5): 808 - 814.
[20] LIANG J, XIAN X, HE X, et al. A novel approach towards medical entity recognition in Chinese clinical text[J]. Journal of healthcare engineering, 2017, 2017.
[21] JENSEN P B, JENSEN L J, Brunak S. Mining electronic health records: towards better research applications and clinical care[J]. Nature reviews genetics, 2012, 13(6): 395 - 405.
[22] 李国垒,陈先来,夏冬,等. 面向临床决策的电子病历文本潜在语义分析[J]. 现代图书情报技术,2016,32(3):50 - 57.
[23] WANG H, ZHANG W, ZENG Q, et al. Extracting important information from Chinese operation notes with natural language processing methods[J]. Journal of biomedical informatics, 2014, 48: 130 - 136.
[24] HE B, DONG B, GUAN Y, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts[J]. Journal of biomedical informatics, 2017, 69: 203 - 217.
[25] 张盈利,夏小玲. 非结构化病理文本的结构化信息抽取方法[J]. 医学信息学杂志,2016,37(4):54 - 58.
[26] 陈德华,冯洁莹,乐嘉锦,等. 中文病理文本的结构化处理方法研究[J]. 计算机科学,2016,43(10):272 - 276.
[27] 丁祥武,张夕华. 医疗领域文本结构化[J]. 计算机工程与设计,2017,38(10):2873 - 2878.
[28] DONG X, CHOWDHURY S, QIAN L, et al. Transfer bi - directional LSTM RNN for named entity recognition in Chinese electronic medical records[C]// Dalian, Liaoning, China: 2017 IEEE 19th International Conference one - Health Networking, Applications and Services (Healthcom). Dalian: IEEE, 2017.
[29] 王鹏远,姬东鸿. 基于多标签 CRF 的疾病名称抽取[J]. 计算机应用研究,2017,34(1):118 - 122.
[30] 侯伟涛,姬东鸿. 基于 Bi-LSTM 的医疗事件识别研究[J]. 计算机应用研究,2018,35(7):1974 - 1977.
[31] BEAN D M, WU H, IQBAL E, et al. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records[J]. Scientific reports, 2017, 7(1): 16416.
[32] ROTMENSCH M, HALPERN Y, TLIMAT A, et al. Learning a health knowledge graph from electronic medical records[J]. Scientific reports, 2017, 7(1): 5994.
[33] 黄梦醒,李梦龙,韩惠蕊. 基于电子病历的实体识别和知识图谱构建的研究[J/OL]. 计算机应用研究:1 - 7[2019 - 03 - 12]. http://kns.cnki.net/kcms/detail/51.1196.TP.20181129.1122.011.html.
[34] CHARIKAR M S. Similarity estimation techniques from rounding algorithms[C]// Montreal, Quebec, Canada: Proceedings of the thirty - fourth annual ACM symposium on Theory of computing. ACM, 2002: 380 - 388.
[35] REZAEIAN N, NOVIKOVA G M. Detecting near-duplicates in Russian documents through using fingerprint algorithm Simhash[J]. Procedia computer science, 2017, 103: 421 - 425.

作者贡献说明:
尚小溥:提出研究思路,设计研究方案,论文起草;
许吴环:采集、清洗和分析数据;
赵红梅:采集、清洗和分析数据;
张润彤:论文最终版本修订;
朱燊:进行实验。

Research on Structure and Knowledge Network Construction of Chinese Ultrasonic Text

Shang Xiaopu[1] Xu Wuhuan[1] Zhao Hongmei[1,2] Zhang Runtong[1] Zhu Shen[1]
[1] Department of Information Management, School of Economic Management, Beijing Jiaotong University, Beijing 100044
[2] Peking University People's Hospital, Beijing 100044

Abstract: [**Purpose/significance**] Ultrasound examination is an important basis for diagnosis, but the major examination data is in the form of text. So, based these data, this paper studies a method that can automatically structure natural language texts and construct knowledge network, which lays the data foundation for further mining clinical knowledge hidden in EMR. [**Method/process**] This paper improved the application of natural language processing technology in ultrasonic, including three main steps: segmentation processing, content location and structured recognition, to realize the segmentation and labeling of ultrasonic text, and on this basis, the ultrasound examination knowledge network was established. [**Result/conclusion**] The test results of real data show that the method for structuring ultrasound texts proposed in this paper has better performance. This method can realize the automatic construction of knowledge network of batch ultrasound texts, and can reflect the potential knowledge of hierarchical relationship and attribute structure of structured content in ultrasonic text.

Keywords: ultrasonic text natural language processing text structuring knowledge network